

IN THE CLAIMS

The following is a complete listing of the claims. This listing replaces all earlier versions and listings of the claims.

Claim 1 (currently amended): A document type definition generating method for generating a document type definition of a structured document containing document elements of a plurality of document element types, wherein each one of the plurality of document element types has a document element name and each document element has a start tag and an end tag, said method comprising:

a physical structure judging step of judging a physical similarity between the document elements in the structured document, wherein the judging of the physical similarity is based on the physical position of the start tag of each document element in the structured document;

a semantic structure judging step of judging a semantic similarity between the document elements by comparing a character string form located between the start tag and the end tag of each of the document elements; and

a document type definition generating step of judging a similarity of the document element tags based on the results obtained in said physical structure judging step and said semantic structure judging step, and generating the document type definition unifying the document element names of similar document elements,

wherein said document type definition generating step includes a redundancy removing step of, when the physical structure and the semantic structure of a plurality of document elements having tags different in element name are judged similar in

said physical structure judging step and said semantic structure judging step, regarding the document elements as being of the same document element type and excluding one document element name from a document type definition generating object based on the judgment results obtained in said physical structure judging step and said semantic structure judging step.

Claim 2 (previously presented): A document type definition generating method according to claim 1, wherein said physical structure judging step includes judging the physical similarity of the document elements based on an indentation or a blank line in the structured document.

Claim 3 (previously presented): A document type definition generating method according to claim 2, wherein, when the physical similarity of the document elements is judged based on the indentation in the structured document in said physical structure judging step, the judging is performed by excluding the indentation which represents a quotation.

Claim 4 (previously presented): A document type definition generating method according to claim 2, wherein, when the physical similarity of the document elements is judged based on the blank line in the structured document in said physical structure judging step, the judging is performed by excluding a determined number of blank lines from the structured document wherein the number of blank lines is determined by constantly skipping one or more blank lines.

Claim 5 (canceled)

Claim 6 (previously presented): A document type definition generating method according to claim 1, wherein said semantic structure judging step includes accessing a semantic information database to judge the semantic similarity of the document element based on a connection of words and phrases in the structured document and word types.

Claim 7 (previously presented): A document type definition generating method according to claim 1, wherein said semantic structure judging step includes judging the semantic similarity of the document element based on a meaning represented by document element tags surrounding the document element.

Claim 8 (canceled)

Claim 9 (currently amended): A document type definition generating method according to claim [[8]] 1, wherein said redundancy removing step includes obtaining similarity degrees concerning agreement degrees of the physical structure and the semantic structure between document elements having tags different in document element name, and regarding the document elements as being of the same type when a general similarity value calculated from the similarity degrees in said redundancy removing step is equal to or greater than a predetermined threshold value.

Claim 10 (previously presented): A document type definition generating method according to claim 1, wherein said document type definition generating step includes a title changing step of, when the physical structure and the semantic structure of a plurality of document elements having document element tags with the same document element name are judged to be different in said physical structure judging step and said semantic structure judging step, regarding the document elements as being of different document element types and changing one document element name based on the judgment results obtained in said physical structure judging step and said semantic structure judging step.

Claim 11 (canceled)

Claim 12 (currently amended): A document type definition generating apparatus for generating a document type definition of a structured document containing document elements of a plurality of document element types, wherein each one of the plurality of document element types has a document element name and each document element has a start tag and an end tag, said apparatus comprising:

physical structure judging means for judging a physical similarity between the document elements in the structured document, wherein the judging of the physical similarity is based on the physical position of the start tag of each document element in the structured document;

semantic structure judging means for judging a semantic similarity between the document elements by comparing a character string form located between the start tag and the end tag of each of the document elements ; and

document type definition generating means for judging a similarity of the document element tags based on the results of said physical structure judging means and said semantic structure judging means, and generating the document type definition unifying the document element names of similar document elements,

wherein said document type definition generating means includes
redundancy removing means for, when the physical structure and the semantic structure of
a plurality of document elements having tags different in element name are judged similar
by said physical structure judging means and said semantic structure judging means,
regarding the document elements as being of the same document element type and
excluding one document element name from a document type definition generating object
based on the judgment results of said physical structure judging means and said semantic
structure judging means.

Claim 13 (previously presented): A document type definition generating apparatus according to claim 12, wherein said physical structure judging means judges the physical similarity of the document elements based on an indentation or a blank line in the structured document.

Claim 14 (previously presented): A document type definition generating apparatus according to claim 13, wherein said physical structure judging means judges the

physical similarity of the document elements based on the indentation by excluding the indentation which represents a quotation.

Claim 15 (previously presented): A document type definition generating apparatus according to claim 13, wherein said physical structure judging means judges the physical similarity of the document elements based on the blank lines by excluding a determined number of blank lines from the structured document wherein the number of blank lines is determined by constantly skipping one or more blank lines.

Claim 16 (canceled)

Claim 17 (previously presented): A document type definition generating apparatus according to claim 12, wherein said semantic structure judging means accesses a semantic information database to judge the semantic similarity of the document element based on a connection of words and phrases in the structured document and word types.

Claim 18 (previously presented): A document type definition generating apparatus according to claim 12, wherein said semantic structure judging means judges the semantic similarity of the document element based on a meaning represented by document element tags surrounding the document element.

Claim 19 (canceled)

Claim 20 (currently amended): A document type definition generating apparatus according to claim [[19]] 12, wherein said redundancy removing means obtains similarity degrees concerning agreement degrees of the physical structure and the semantic structure between document elements having tags different in element name, and regards the document elements as being of the same document element type when a general similarity value calculated from the similarity degrees by said redundancy removing means is equal to or greater than a predetermined threshold value.

Claim 21 (previously presented): A document type definition generating apparatus according to claim 12, wherein said document type definition generating means includes title changing means for, when the physical structure and the semantic structure of a plurality of document elements having document element tags with the same element name are judged to be different by said physical structure judging means and said semantic structure judging means, regarding the document elements as being of different document element types and changing one document element name based on the judgment results of said physical structure judging means and said semantic structure judging means.

Claim 22 (canceled)

Claim 23 (currently amended): A computer-readable storage medium storing a program for controlling a computer to execute a document type definition generation method for generating document type definition of a structured document containing document elements of a plurality of document element types, wherein each one

of the plurality of document element types has a document element name and each document element has a start tag and an end tag, said program comprising:

code for a physical structure judging step of judging a physical similarity between the document elements in the structured document, wherein the judging of the physical similarity is based on the physical position of the start tag of each document element in the structured document;

code for a semantic structure judging step of judging a semantic similarity between the document elements by comparing a character string form located between the start tag and the end tag of each of the document elements; and

code for a document type definition generating step of judging a similarity of the document element tags based on the results obtained by said physical structure judging code and said semantic structure judging code, and generating the document type definition unifying the document element names of similar document elements,

wherein said code for a document type definition generating step includes code for a redundancy removing step of, when the physical structure and the semantic structure of a plurality of document elements having tags different in element name are judged similar by said code for a physical structure judging step and said code for a semantic structure judging step, regarding the document elements as being of the same document element type and excluding one document element name from a document type definition generating object based on the judgment results obtained by said code for a physical structure judging step and said code for a semantic structure judging step.

Claim 24 (currently amended): A processing method for processing a structured document containing document elements of a plurality of document element types, wherein each one of the plurality of document element types has a document element name and each document element has a start tag and an end tag, said method comprising:

an input step of inputting the structured document; [[and]]
a judging step of judging the semantic similarity between the document elements by comparing a character string form located between the start tag and the end tag of each document element; and

a processing step of regarding the document elements as the same document element type, and executing a predetermined process based on the document elements being regarded as the same document element type when the semantic structures of the document elements are judged similar in said judging step.

Claim 25 (currently amended): A processing method for processing a structured document containing document elements of a plurality of document element types, wherein each one of the plurality of document element types has a document element name and each document element has a start tag and an end tag, said method comprising:

an input step of inputting the structured document; [[and]]
a judging step of judging the physical similarity between the document elements based on a physical similarity of the document elements according to positions of each document element start tag in the structured document; and

a processing step of regarding the document elements as the same document element type, and executing a predetermined process based on the document

elements being regarded as the same document element type when the positions of each start tag in the structured document are judged similar in said judging step.

Claim 26 (currently amended): A processing apparatus for processing the similarity of document elements in a structured document containing document elements of a plurality of document element types, wherein each one of the plurality of document element types has a document element name and each document element has a start tag and an end tag, said apparatus comprising:

an input device for inputting the structured document; and
a judging device for judging the semantic similarity between the document elements by comparing a character string form located between the start tag and the end tag of each document element,

wherein said judging device regards the document elements as the same document element type, and executes a predetermined process on the document elements being regarded as the same document element type when the semantic structures of the document elements are judged similar.

Claim 27 (currently amended): A processing apparatus for processing the similarity of document elements in a structured document containing document elements of a plurality of document element types, wherein each one of the plurality of document element types has a document element name and each document element has a start tag and an end tag, said apparatus comprising:

an input device for inputting the structured document; and

a judging device for judging the physical similarity between the document elements based on a physical similarity of the document elements according to positions of each document element start tag in the structured document,

wherein said judging device regards the document elements as the same document element type, and executes a predetermined process based on the document elements being regarded as the same document element type when the positions of each start tag in the structured document are judged similar.

Claim 28 (new): A method according to claim 24, wherein said judging step includes accessing a semantic information database to judge the semantic similarity of the document elements based on a connection of words and phrases in the structured document and word types.

Claim 29 (new): A method according to claim 25, wherein said judging step includes judging the physical similarity of the document elements based on an indentation or blank line in the structured document.

Claim 30 (new): A method according to claim 25, wherein the judging is performed in said judging step by excluding the indentation which represents a quotation when the physical similarity of the document elements is judged based on the indentation.

Claim 31 (new): An apparatus according to claim 26, wherein said judging device accesses a semantic information database to judge the semantic similarity of the

document elements based on a connection of words and phrases in the structured document and word types.

Claim 32 (new): An apparatus according to claim 27, wherein said judging device judges the physical similarity of the document elements based on an indentation or a blank line in the structured document.

Claim 33 (new): An apparatus according to claim 27, wherein the judging is performed by said judging device by excluding the indentation which represents a quotation when the physical similarity of the document elements is judged based on the indentation.